Random Variables

- We've been working with random variables all semester, we just haven't called them that.
- A *Random variables* is just a numerical variable in statistics, i.e. a random outcome that is quantitative (numerical).

Example: *Sally, Bill, George and Bettina make up a small engineering team. Sally makes $75,000 a year, Bill makes $35,000, George makes $100,000 and Bettina makes $75,000. Consider two random sampling procedures:*

- *A procedure that doesn't generate a random variable:* For the procedure "select a person at random from the team", the sample space is

$$S = \{\text{Sally, Bill, George, Bettina}\}$$

and if the outcomes are equally likely then each has probability 1/4 of occurring.

- *A procedure that does generate a random variable:* For the procedure "select a person at random from the team and check their salary" the sample space is

$$S = \{35000,\ 75000,\ 100000\}$$

(only three outcomes since one salary is repeated).

In the language of random variables we would say:

"Let X be a random variable with values 35000, 75000, and 100000. Note that these values are no longer *equally likely. Rather,*

Outcome	*Probability*
$35,000	*1/4*
$75,000	*1/2**
$100,000	*1/4*

*Two people have this salary.

These probabilities are called the distribution *of X."*

Expected Value (Central Location)

In our salary example, suppose you repeat 100 times the procedure of selecting a person from the team at random and checking their salary. What would the sample average of your 100 observations be?

Since you expect to see each person about 25 times (1/4 of the 100 trials), you'd expect your sample average to be about

$$\begin{aligned}\overline{X} &\approx \frac{(25)(35000)+(50)(75000)+(25)(100000)}{100} \\ &= (1/4)(35000)+(1/2)(75000)+(1/4)(100000) \\ &= \$71,250.\end{aligned}$$

Note that $71,250 is the sum of each salary times its probability. This forms the basis of the definition of the expected value *of a random variable.*

The *expected value* or *mean* of the random variable X with values x is

$$\begin{aligned}\mu &= \text{Sum of values times probabilities} \\ &= \sum x \cdot P(X = x)\end{aligned}$$

Standard Deviation (Spread or variability)

You can use the same idea to predict the sample standard deviation in repeated independent trials:

The *standard deviation* (SD) of the random variable X with values x is

$$\begin{aligned} \sigma &= \sqrt{\text{Sum of squared deviations times probabilities}} \\ &= \sqrt{\sum (x-\mu)^2 P(X=x)} \end{aligned}$$

In the salary example above, the variance of the salary random variable X is

$$\begin{aligned} (1/4)(35000-71250)^2 &+ (1/2)(75000-71250)^2 \\ &+ (1/4)(100000-71250)^2 \\ &= 542,187,500 \end{aligned}$$

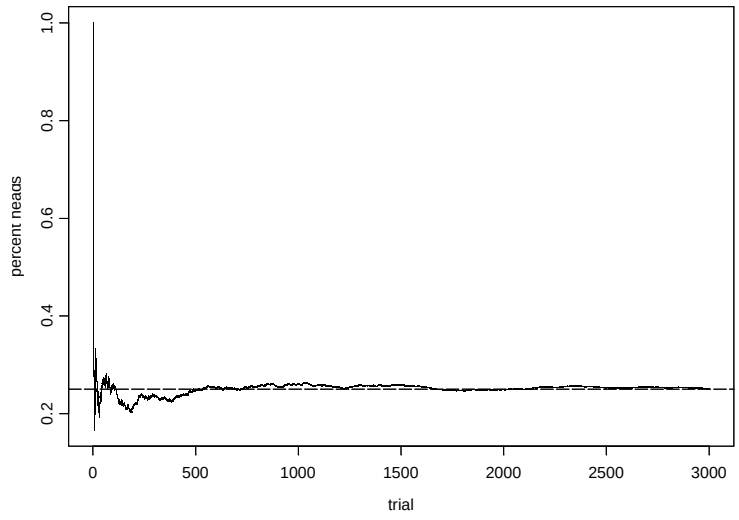
and the SD of X is

$$\sigma = \sqrt{542,187,500} = \$23,285.$$

This is a very large SD, but of course the range of salaries from $35,000 to $100,000 is large too.

Law of Large Numbers

If a random procedure with numerical outcomes is repeated many times independently, the mean value of the actually-observed outcomes approaches the true mean or expected value.



- The picture shows the law of large numbers for an "unfair coin" with probability of heads = 0.25.
- The same principle applies to averages from *any* sample where the members of the sample were observed independently.
- The larger the standard deviation (SD), the longer we have to wait for the law of large numbers to "kick in".

Gambling

A Simple Lottery

In a simple lottery, 100,000 tickets are sold for $1.00 each and the following prizes are guaranteed:

18	$400 prizes
120	$50 prizes
270	$40 prizes

The expected payout on a $0.50 ticket is

$$\frac{18}{100000}\cdot(\$400)+\frac{120}{100000}\cdot(\$50)+\frac{270}{100000}\cdot(\$40)=\$0.24$$

so the state only pays out 48% (= .24/.50) of the money wagered in the lottery.

From `http://palottery.com/`*: "The daily number"*

- Pick any three digit number
- Choose a bet from \$0.50 to \$5.00
- Perfect match pays 500 to 1, so if you bet \$1.00 a perfect match pays \$500.
- Prob of an exact match on three digits:

$$\frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} = 0.001, \text{ or 1 in 1000.}$$

- For a \$1.00 bet, the expected payout is

$$\$500 \cdot 0.001 + \$0.00 \cdot 0.999 = \$0.50.$$

- PA is paying out exactly 50% of the wager in this game.
- It is very typical for a state lottery to pay out about 1/2 of what is wagered.

(Moore: 15% $\rightarrow$ advertising; 35% $\rightarrow$ state coffers; Casinos pay more (85%–95%)).

From `http://palottery.com/`: *"Super 6 Lotto"*

[Approximate analysis]

- A "play" costs $1.00 (three sets of 6 numbers).
- State skims 48 cents off the top and puts 52 cents in "Prize pool".
- Player picks 6 numbers from 1 to 69. Winning odds below (odds $1 : x \leftrightarrow$ probability $1/(1+x)$).

Match	Odds	Percent of Pool Won	Winnings on $1.00
6 of 6	1:39,959,158	76.0%	(.760)($0.52) = $0.3952
5 of 6	1:105,715	8.0%	(.080)($0.52) = $0.0416
4 of 6	1:1,364	7.5%	(.075)($0.52) = $0.0390
3 of 6	1:51	8.5%	(.085)($0.52) = $0.0442

- Expected return on $1.00 wager:

$$\$0.3952 \cdot (1/39,959,159) \; + \; \$0.0416 \cdot (1/105,716)$$
$$+ \; \$0.0390 \cdot (1/1,365) \; + \; \$0.0442 \cdot (1/52) = \$0.000863.$$

- So the State expects to pay you back less than 1/10 of a penny on each dollar you spend on the lottery.
- *What about those big prizes?* The betting system is paramutuel (everyones' wagers increase the pot) and unclaimed money is rolled over to a new play.
- So entering lotto after some rollovers can eventually make the bet fair to you (but not to last week's losers).

Disentangling Probability by Simulation

The Birthday Problem

What's the probability of two [or more] people in a group having the same birthday?

```
MTB > set c1
DATA> 1:365
DATA> end.
MTB > Sample 10 C1 c2;
SUBC>   Replace.
MTB > Tally C2;
SUBC>   Counts.

      C2  Count
      42      1
     154      1
     163      1
     168      1
     206      1
     212      1
     224      1
     250      1
     309      1
     313      1
      N=     10
```

```
MTB > Sample 10 C1 c2,
SUBC>   Replace.
MTB > Tally C2;
SUBC>   Counts.

      C2  Count
      55      1
      74      1
     133      2
     180      1
     192      1
     241      1
     244      1
     257      1
     346      1
      N=     10
```

Do this many times (10,000!) on the computer:

Number of people	Prob of at least one match
10	0.1209
20	0.4075
25	0.5658
80	1.0000

Binomial

A random variable that we encounter repeatedly in opinion surveys, marketing surveys, quality and performances studies, etc., is the *binomial* random variable.
If an experiment is analogous to counting the number X of heads in n independent flips of a loaded coin, then

- *The* probability of Heads *on each flip is denoted p.*
- *X (the number of heads) is called a* binomial random variable.
- *$\hat{p} = X/n$ (the fraction of heads) is called a* binomial proportion.

We have already seen the formulas for the mean μ and standard deviation σ of a binomial random variable:

$$\mu_X = np \quad \sigma_X = \sqrt{np(1-p)}$$

$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

We can also calculate—exactly—the probability of a certain number of successful outcomes from a binomial distribution:

$$P[X = k] = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{(for } k = 0, 1, 2, \ldots n\text{)}$$

The symbol "$\binom{n}{k}$" stands for "n choose k", the binomial coefficient from high-school algebra:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

The symbol "$n!$" is read "n factorial".

$n!$	**Meaning**	**Value**
$0!$	(special case)	1
$1!$	1	1
$2!$	$2 \cdot 1$	2
$3!$	$3 \cdot 2 \cdot 1$	6
$4!$	$4 \cdot 3 \cdot 2 \cdot 1$	24
$5!$	$5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$	120
$6!$	$6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$	720
$7!$	$7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$	5040
$8!$	$8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$	40320

Where does it come from?

- Imagine a loaded coin with P[heads] $= 0.25$. We want the probability of *exactly three heads in five independent tosses*.
- That's three heads, two tails:

  $$(0.25)(0.25)(0.25)(0.75)(0.75) = (0.25)^3(0.75)^2$$

  But this looks like only the case HHHTT…
- What about HTHTH? That's

  $$(0.25)(0.75)(0.25)(0.75)(0.25) = (0.25)^3(0.75)^2$$

  again, same answer! We should add these up because they are both possible outcomes.
- How many other possibilities are there? $\binom{5}{3}$. Adding all these we get

  $$\binom{5}{3}(0.25)^3(0.75)^2$$

- This is just the formula

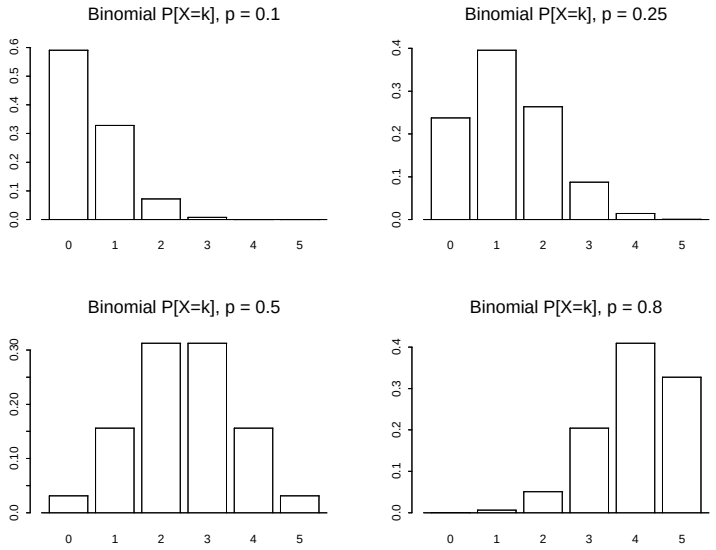
  $$P[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

  with $n = 5$, $k = 3$ $p = 0.25$.

So…, what is the probability of three heads in five tosses of that loaded coin?

$$\begin{aligned} P[X=k] &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \binom{5}{3} (0.25)^3 (1-0.25)^{5-3} \\ &= \frac{5!}{3! \cdot 2!} (0.25)^3 (0.75)^2 \\ &= \frac{120}{(6)(2)} (0.25)^3 (0.75)^2 \\ &= (10) \cdot (0.008789062) \\ &= 0.08789062 \end{aligned}$$

We can look at all the probabilities in an idealized histogram:



The 68–95–99.7 doesn't work as well for skewed distributions as for symmetric ones.